# Two-Step Sequential Sampling

Moors, J.J.A.; Strijbosch, L.W.G.

*Publication date:*
2000

Center
for
Economic Research

No. 2000-39

**TWO STEP SEQUENTIAL SAMPLING**

By J.J.A. Moors and L.W.G. Strijbosch

April 2000

ISSN 0924-7815

# TWO-STEP SEQUENTIAL SAMPLING

J.J.A. Moors & L.W.G. Strijbosch

**Abstract**

Deciding upon the optimal sample size in advance is a difficult problem in general. Often, the investigator regrets not having drawn a larger sample; in many cases additional observations are done. This implies that the actual sample size is no longer deterministic; hence, even if all sample elements are drawn at random, the final sample is *not* a simple random sample. Although this fact is widely recognized, its consequences are often grossly underrated in our view. Too often, these consequences are ignored: the usual statistical procedures are still applied.

This paper shows in detail the dangers of applying standard techniques to extended samples. To allow theoretical derivations only some elementary situations are considered. More precisely, the following features hold throughout the paper:

- the population variable of interest is normally distributed,
- estimation concerns population mean and variance,
- all sample elements are drawn at random, with replacement,
- only standard estimators, like sample mean and sample variance, will be considered.

Nevertheless, the results are rather disturbing: standard estimators have sizable biases, their variances are (much) larger than usual, and standard confidence intervals do not have the prescribed confidence level any more.

Crucial is of course the criterion used to decide whether or not to extend the original sample. Four criteria are applied. In the first three cases, an independent event, the observed sample mean and the observed sample variance, respectively, determine whether or not to double the original sample size. The fourth criterion compares the variances observed in two independent samples; the sample with the highest variance is extended. Only in this fourth case the size of the extension is a random variable. Note that a given criterion is used only once: after the original observations the final sample size is determined; hence the title of our paper.

# 1 Introduction

This paper considers the following type of two-step sampling procedures. In the first step a random sample of fixed size is drawn and observed; depending on the outcomes an additional sample may be drawn. The size of this second step sample may be fixed or stochastic. In both cases, combining the observations of both steps leads to a stochastic final sample size, leading to interesting problems in drawing conclusions about population parameters.

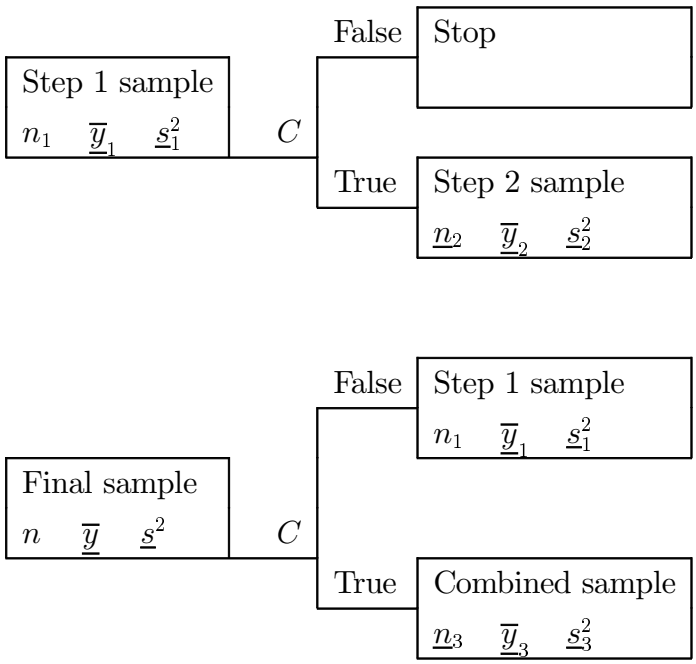
This general situation arises for example if the accuracy of first step estimates is not satisfactory: the researcher will be tempted to do additional observations. Or the first step can be considered as a pilot survey: the final sample size is determined on the base of the estimated standard deviation in the population.

The usual inferential procedure in these cases is to condition on the final sample size, that is to use the standard methods for finding estimates and their accuracy, based on either the first step observations only, or on the combined observations of both steps. It will be shown, however, that this approach is (very) questionable - depending of course on the criterion $C$ for doing step 2.

Attention will be focussed on estimating the population mean $\mu$; to measure accuracy, estimates of the population variance $\sigma^2$ will be needed as well. Furthermore, all theoretical results concern normal distributions $N(\mu, \sigma^2)$; these results are checked by means of simulations.

Figure 1.1 summarizes the two-step sampling procedure outlined above, as well as some necessary notation. The stochastic variables $\underline{n}, \underline{\overline{y}}$ and $\underline{s}^2$ indicate size, mean and variance of the final sample; the indices 1, 2 and 3 indicate step 1, step 2 and the combination of both steps, respectively. The upper scheme shows the (possible) extension of the step 1 sample; the lower scheme gives the details of the final sample.

**Figure 1.1**. General two-step sampling procedure.




As stated before, crucial is the criterion for extending the sample with step 2. The following four extension criteria will be considered; $c$ indicates a given constant.

$$C_1 = \{\text{event, independent of step 1 sample}\}$$

$$C_2 = \{\underline{\overline{y}}_1 > c\}$$

$$C_3 = \{\underline{s}_1^2 > c\}$$

$$C_4 = \{\underline{s}_1^2 > \text{ variance in another sample}\}$$

These cases will be treated in the subsequent sections. $C_1$ illustrates the consequences of a stochastic sample size in general. Extension criteria $C_2$ and $C_3$ are related to the outcomes of step 1. An application of $C_2$ is found in the field of hypothesis testing in auditing; $C_3$ concerns the familiar situation that the sample is extended because of insufficient accuracy. Criterion $C_4$, finally, has to do with two-sample problems, for example the situation where two normal means have to be estimated with approximately equal accuracy.

In all cases, the expectation $E(\underline{\overline{y}})$ and $\mathrm{Var}(\underline{\overline{y}})$ of the usual estimator $\underline{\overline{y}}$ for $\mu$ will be derived. Besides, the expectations of the estimators $\underline{s}^2$ for $\sigma^2$ and $\underline{s}^2/\underline{n}$ for $\mathrm{Var}(\underline{\overline{y}})$ will

be calculated. For $C_2$, hypothesis testing with respect to $\mu$ will be discussed. In the first three cases, the step 2 sample size is taken fixed and equal to the size of step 1.

# 2 Independent extension criterion

In this section, some independent random experiment, like rolling a die determines whether or not an additional sample of size $\underline{n}_2 = n_1$ will be drawn. If so, the final sample mean is the average of the means obtained in both steps separately. Hence

$$\underline{n} = \begin{cases} n_1 & \text{if} \quad C_1' \\ 2n_1 & \text{if} \quad C_1 \end{cases}$$

$$\overline{\underline{y}} = \begin{cases} \overline{\underline{y}}_1 & \text{if} \quad C_1' \\ (\overline{\underline{y}}_1 + \overline{\underline{y}}_2)/2 & \text{if} \quad C_1 \end{cases} \tag{2.1}$$

Since $\overline{\underline{y}}$ is unbiased for $\mu$ conditionally, both under $C_1$ and its complement $C_1'$, with variance $\sigma^2/n_1$ and $\sigma^2/(2n_1)$, respectively, $\overline{\underline{y}}$ is unbiased unconditionally as well, with variance

$$\mathrm{Var}(\overline{\underline{y}}) = \sigma^2 E(1/\underline{n}) \tag{2.2}$$

Formally, this result can be found from the familiar formulae

$$\begin{cases} E(\underline{x}) &= E[E(\underline{x}|\underline{y})] \\ \mathrm{Var}(\underline{x}) &= E[\mathrm{Var}(\underline{x}|\underline{y})] + \mathrm{Var}[E(\underline{x}|\underline{y})] \end{cases} \tag{2.3}$$

holding for any pair of random variables $(\underline{x}, \underline{y})$.

Similarly, the straightforward estimators

$$\underline{s}^2 = \begin{cases} \underline{s}_1^2 & \text{if} \quad C_1' \\ \underline{s}_3^2 & \text{if} \quad C_1 \end{cases} \tag{2.4}$$

$$\underline{\mathrm{var}}(\overline{\underline{y}}) = \begin{cases} \underline{s}_1^2/n_1 & \text{if} \quad C_1' \\ \underline{s}_3^2/(2n_1) & \text{if} \quad C_1 \end{cases} \tag{2.5}$$

are unbiased for $\sigma^2$ and $\mathrm{Var}(\overline{\underline{y}})$, respectively.

Denoting $p = P(C_1)$, (2.1) implies

$$E(\underline{n}) = n_1(1+p) \qquad E(1/\underline{n}) = (1 - p/2)/n_1$$

A sample of *fixed* size $E(\underline{n})$ - neglecting rounding - would have led to an estimated population mean with variance

$$\text{Var}^{\,*}(\overline{\underline{y}}) = \sigma^2/[n_1(1+p)]$$

Hence, the loss in accuracy due to the stochastic sample size is given by

$$\text{Var}(\overline{\underline{y}})/\,\text{Var}^{\,*}(\overline{\underline{y}}) - 1 = p(1-p)/2$$

So, at the same expected costs, the variance is up to 12.5% higher than necessary.

Note that up to now the normality assumption was irrelevant. However, in case of normality, the straightforward confidence interval, having the limits

$$\overline{\underline{y}} \pm t_{\underline{n};\alpha/2}\sqrt{\underline{\text{var}}(\overline{\underline{y}})} = \begin{cases} \overline{\underline{y}}_1 \pm t_{n_1;\alpha/2}\sqrt{\underline{\text{var}}(\overline{\underline{y}}_1)} & \text{if} \quad C_1' \\ \overline{\underline{y}}_3 \pm t_{2n_1;\alpha/2}\sqrt{\underline{\text{var}}(\overline{\underline{y}}_3)} & \text{if} \quad C_1 \end{cases}$$

indeed has confidence level $1-\alpha$, since both conditional probabilities of containing $\mu$ equal $1-\alpha$.

# 3 Extension based on sample mean

Now, a constant $c$ is chosen; if the step 1 sample mean exceeds $c$, a second sample of size $n_1$ is drawn. With obvious adaptation expression (2.1) still holds; standardization and introduction of

$$d = \frac{c-\mu}{\sigma/\sqrt{n_1}}$$

gives

$$\frac{\overline{\underline{y}}-\mu}{\sigma/\sqrt{n_1}} = \begin{cases} \underline{z}_1 & \text{if} \quad \{\underline{z}_1 \le d\} = C_2' \\ (\underline{z}_1+\underline{z}_2)/2 & \text{if} \quad \{\underline{z}_1 > d\} = C_2 \end{cases} \tag{3.1}$$

with $\underline{z}_1, \underline{z}_2$ independent standard normal variables. Consequently,

$$E\left[\frac{\overline{\underline{y}}-\mu}{\sigma/\sqrt{n_1}}\right] = E(\underline{z}_1|C_2')P(C_2') + E[(\underline{z}_1+\underline{z}_2)/2|C_2]P(C_2)$$

Using

$$0 = E(\underline{z}_1) = E(\underline{z}_1|C_2')P(C_2') + E(\underline{z}_1|C_2)P(C_2)$$

this can be simplified to

$$\begin{aligned} E\left[\sqrt{n_1}(\underline{\overline{y}}-\mu)/\sigma\right] &= E[(\underline{z}_2-\underline{z}_1)/2|C_2]P(C_2) \\ &= -\tfrac{1}{2}\textstyle\int_d^\infty z_1\varphi(z_1)dz_1 = -\tfrac{1}{2}[-\varphi(z_1))]_d^\infty = -\tfrac{1}{2}\varphi(d) \end{aligned}$$

since the conditional expectation of $\underline{z}_2$ equals 0. Hence, the final estimator $\underline{\overline{y}}$ has expectation

$$E(\underline{\overline{y}}) = \mu - \frac{\sigma}{2\sqrt{n_1}}\varphi\left(\frac{c-\mu}{\sigma/\sqrt{n_1}}\right) \tag{3.2}$$

Logically, $\underline{\overline{y}}$ is negatively biased.

Similarly,

$$\begin{aligned} E\left[(\sqrt{n_1}(\underline{\overline{y}}-\mu)/\sigma)^2\right] &= 1+E[((\underline{z}_1+\underline{z}_2)^2/4-\underline{z}_1^2|C_2)]P(C_2) \\ &= 1-\tfrac{3}{4}E(\underline{z}_1^2|C_2)P(C_2)+E\left(\underline{z}_1\underline{z}_2/2+\underline{z}_2^2/4|C_2\right)P(C_2) \\ &= 1-\tfrac{3}{4}\textstyle\int_d^\infty z_1^2\varphi(z_1)dz_1+\tfrac{1}{4}P(C_2) \\ &= 1-\tfrac{3}{4}[\Phi(z_1)-z_1\varphi(z_1)]_d^\infty+\tfrac{1}{4}[1-\Phi(d)] \end{aligned}$$

so that the mean squared error of $\underline{\overline{y}}$ equals

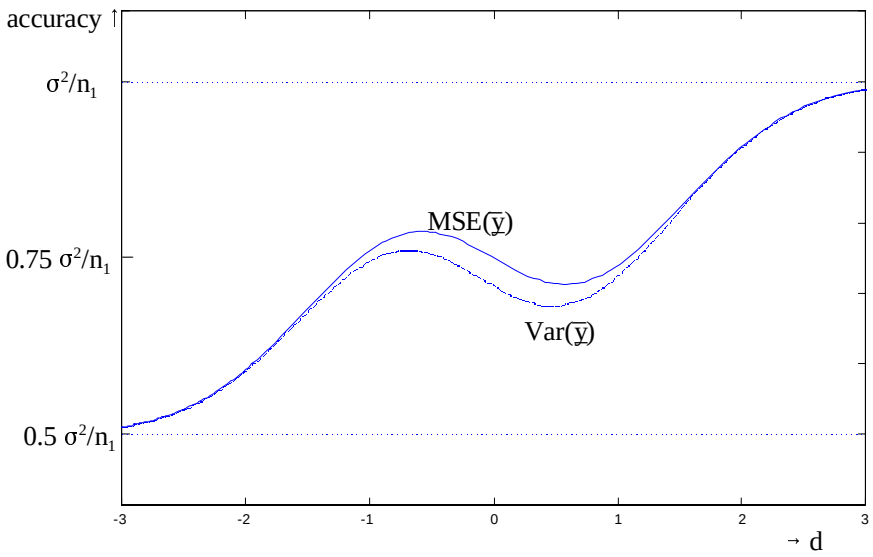
$$\text{MSE }(\underline{\overline{y}}) = \frac{\sigma^2}{2n_1}\left[1+\Phi(d)-\frac{3}{2}d\varphi(d)\right] \tag{3.3}$$

Subtracting the squared bias gives

$$\text{Var }(\underline{\overline{y}}) = \frac{\sigma^2}{2n_1}\left[1+\Phi(d)-\frac{3}{2}d\varphi(d)-\frac{1}{2}\varphi^2(d)\right]$$

Figure 3.1 shows these two accuracy measures as a function of $d=\sqrt{n_1}(c-\mu)/\sigma$.

**Figure 3.1**. Var $(\underline{\overline{y}})$ and MSE $(\underline{\overline{y}})$ using $C_2$.




Note that MSE $(\underline{\overline{y}})$ has local extrema for $d = \pm\sqrt{1/3}$ and Var $(\underline{\overline{y}})$ for $d = -0.6915$ and $d = 0.4704$. The asymptotic values agree with intuition. The decrease of both accuracy measures for $d \approx 0$ can be explained from the fact that large $\underline{\overline{y}}_1$-values are corrected downwards by the additional observations.

The expected sample size equals

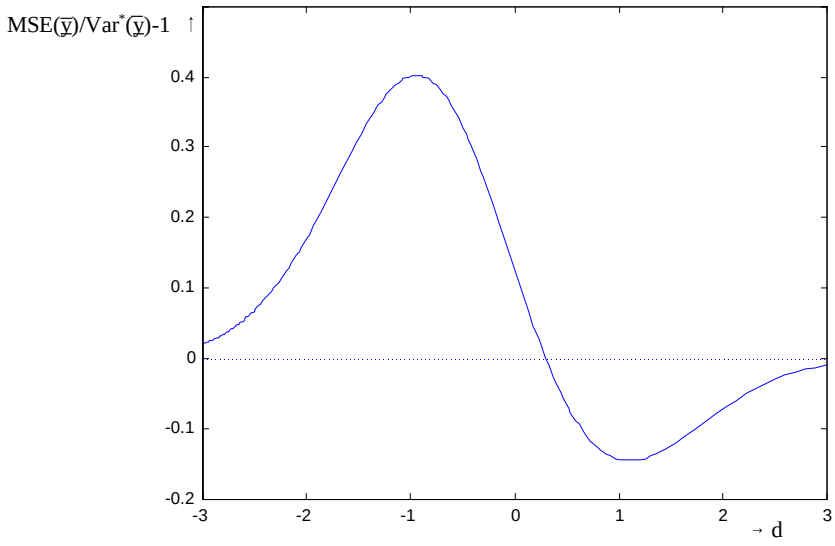
$$E(\underline{n}) = n_1[2 - \Phi(d)]$$

Again denoting with Var $^*(\underline{\overline{y}})$ the variance of a single sample of this fixed size, the loss of accuracy due to the stochastic sample size is:

$$\text{MSE } (\underline{\overline{y}}) / \text{ Var }^*(\underline{\overline{y}}) - 1 = \frac{1}{2}\Phi(d)[1 - \Phi(d)] - \frac{3}{4}d\varphi(d)[2 - \Phi(d)]$$

Compare Figure 3.2.

**Figure 3.2**. Accuracy loss from two-step procedure using $C_2$.




The loss is 0 for $d = 0.298$; extrema are 0.402 for $d = -0.945$ and $-0.1455$ for $d = 1.104$.

Using

$$\sigma^2 = E(\underline{s}_1^2) = E(\underline{s}_1^2|C_2')P(C_2') + E(\underline{s}_1^2|C_2)P(C_2)$$

the expectation of $\underline{s}^2$ in (2.4) can be written as

$$E(\underline{s}^2) = \sigma^2 + E(\underline{s}_3^2 - \underline{s}_1^2|C_2)P(C_2) \tag{3.4}$$

Now, $\underline{s}_3^2$ can be expressed standardly in the separate outcomes of the step 1 and step 2 samples:

$$(2n_1 - 1)\underline{s}_3^2 = (n_1 - 1)\underline{s}_1^2 + (n_1 - 1)\underline{s}_2^2 + \frac{n_1}{2}(\overline{\underline{y}}_1 - \overline{\underline{y}}_2)^2$$

so that

$$\begin{aligned}
(2n_1 - 1)\left[E(\underline{s}^2) - \sigma^2\right] &= E\left[(n_1 - 1)\underline{s}_2^2 - n_1\underline{s}_1^2 + \tfrac{1}{2}n_1(\overline{\underline{y}}_1 - \overline{\underline{y}}_2)^2|C_2\right]P(C_2) \\
&= -\sigma^2 P(C_2) + \tfrac{1}{2}E\left[n_1(\overline{\underline{y}}_1 - \overline{\underline{y}}_2)^2|\overline{\underline{y}}_1 > c\right]P(C_2) \\
&= -\sigma^2 P(C_2) + \tfrac{1}{2}\sigma^2 E\left[(\underline{z}_1 - \underline{z}_2)^2|\underline{z}_1 > d\right]P(\underline{z}_1 > d) \\
&= \tfrac{1}{2}\sigma^2 d\varphi(d)
\end{aligned}$$

Here, $\underline{z}_i$ is again the standardized $\underline{\overline{y}}_i$ $(i = 1, 2)$; further, compare the derivation of (3.3). Hence, the variance estimator $\underline{s}^2$ is biased:

$$E(\underline{s}^2) = \sigma^2 \left[1 + \frac{1}{2(2n_1 - 1)} d\varphi(d)\right] \tag{3.5}$$

Similarly, (2.5) gives

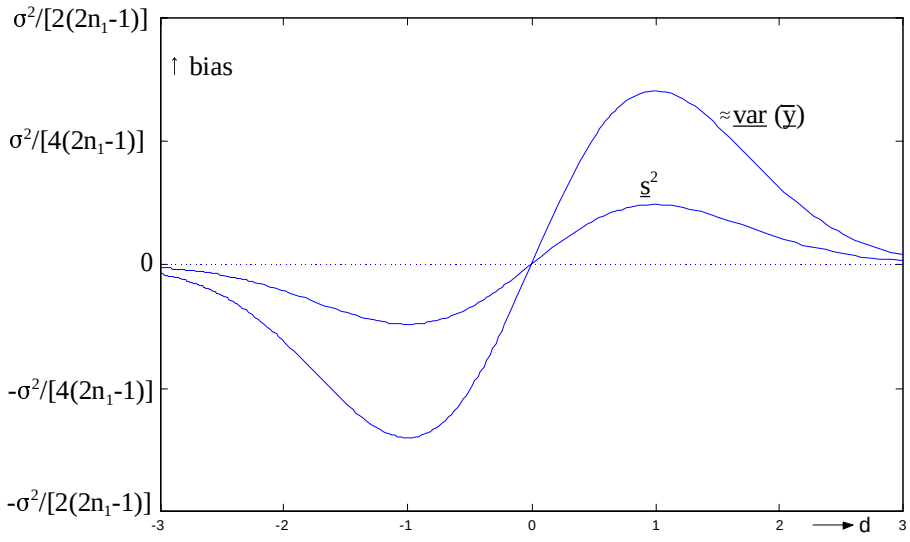
$$\begin{aligned} E[\underline{\text{var}}(\underline{\overline{y}})] &= \frac{\sigma^2}{n_1} + \tfrac{1}{2}E[\underline{s}_3^2 - 2\underline{s}_1^2 | C_2]P(C_2)/n_1 \\ &= \tfrac{1}{2}\sigma^2 \left[1 + \Phi(d) + \frac{1}{2(2n_1 - 1)} d\varphi(d)\right] / n_1 \end{aligned}$$

Comparison with (3.4) shows

$$E[\underline{\text{var}}(\underline{\overline{y}})] = \text{MSE } (\underline{\overline{y}}) + \frac{3n_1 - 1}{2n_1(2n_1 - 1)} \sigma^2 d\varphi(d)$$

The bias $B[\underline{\text{var}}\,(\underline{\overline{y}})]$ therefore is approximately equal to $3B(\underline{s}^2)$; Figure 3.3 shows both biases.

**Figure 3.3**. Biases of $\underline{s}^2$ and $\underline{\text{var}}\,(\underline{\overline{y}})$ using $C_2$.




Both estimators are unbiased for $d = 0$ or $d \to \pm\infty$; local extrema occur for $d = \pm 1$.

For worst-case scenarios, the extreme value of the relative bias (notation: ERB) is of importance. It is easy to see that

$$\text{ERB } (\underline{\overline{y}}) = -\frac{0.199\sigma/\mu}{\sqrt{n_1}}$$

while for not too small $n_1$

$$\text{ERB } (\underline{s}^2) \approx \frac{0.0605}{n_1} \qquad \text{ERB}[\underline{\text{var}}\,(\overline{\underline{y}})] \approx 0.181 \tag{3.6}$$

Note that these worst cases occur for $d = 1$; for fixed $c$, the situation is much less serious: even the last bias tends to 0 for $n_1 \to \infty$.

These theoretical expressions were checked by means of simulation. Following the scheme in Figure 1.1, $N$ final samples were generated from a standard normal distribution; for each final sample the values of $\overline{y}$, $s^2$ and var($\overline{y}$) were determined. All combinations of $d \in \{0.5, 1.0, 1.5\}$ and $n_1 \in \{4, 9, 16, 25\}$ were used. The number of final samples were chosen such that $n_1 * N = 1,000,000$ (or approximately). For efficiency reasons, for any $d$ the same observations were used for all values of $n_1$; so, in total nearly 2 million standard normal observations were drawn.

The average values of the statistics $\overline{y}$, $s^2$ and var($\overline{\underline{y}}$) were calculated over the $N$ replications, as well as their biases. In Appendix B these empirical biases are compared with the theoretical values; the overall agreement is good.

Since the variance estimator $\underline{\text{var}}(\overline{\underline{y}})$ is the most interesting statistic, histograms of its $N$ individual values are presented in Figures 3.4 and 3.5. In each diagram the mean of all var($\overline{\underline{y}}$) values is displayed as well as the corresponding theoretical values. The bars in the diagrams are 'stacked': the lower part represents the final samples containing $n_1$ observations ($C_2$ is false) and the upper part $2n_1$ ($C_2$ is true).

**Figure 3.4.** Empirical distribution of var($\overline{\underline{y}}$) using $C_2$ ($n_1 = 9$; $N$ =110,000).

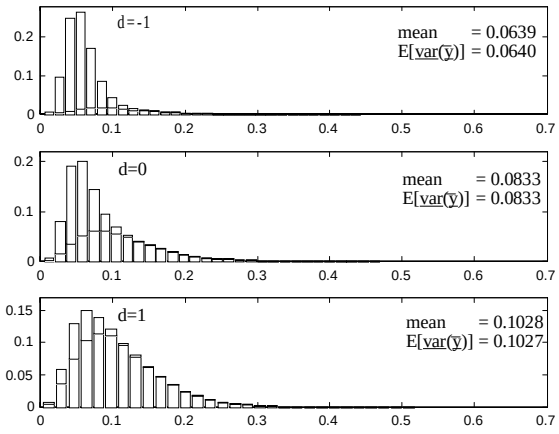
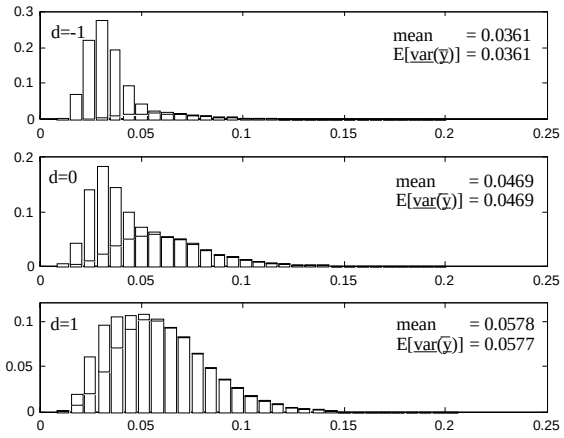


**Figure 3.5.** Empirical distribution of var($\underline{\overline{y}}$) using $C_2$ ($n_1 = 16$; $N$ =62,500).




An example of two-step sequential sampling with extension criterion $C_2$ can be found in auditing. In order to determine whether a certain auditing population can be approved of, the following testing problem may be considered:

$$H_0 : \mu \geq m, \quad H_a : \mu < m$$

Here, $\mu$ is the mean error in the population values and $m$ is the so-called materiality: the maximum mean error that can be allowed according to the auditor. Based on a random sample of $n_1$ values, the standard test procedure reads

$$\begin{cases} \underline{\overline{y}}_1 < m - z_\alpha \sigma/\sqrt{n_1} & \Longrightarrow \quad \text{reject } H_0 \\ \underline{\overline{y}}_1 \geq m - z_\alpha \sigma/\sqrt{n_1} & \Longrightarrow \quad \text{do not reject } H_0 \end{cases} \tag{3.7}$$

(where a normal distribution with known $\sigma$ has been assumed). The latter part of this decision rule is, however, often interpreted as "the auditing efforts were not sufficient" rather than "the population can not be approved of". So, in practice, an additional sample will often be drawn. If this second sample is of size $n_1$ as well, (3.1) is obtained, now with

$$d = \frac{m - \mu}{\sigma/\sqrt{n_1}} - z_\alpha$$

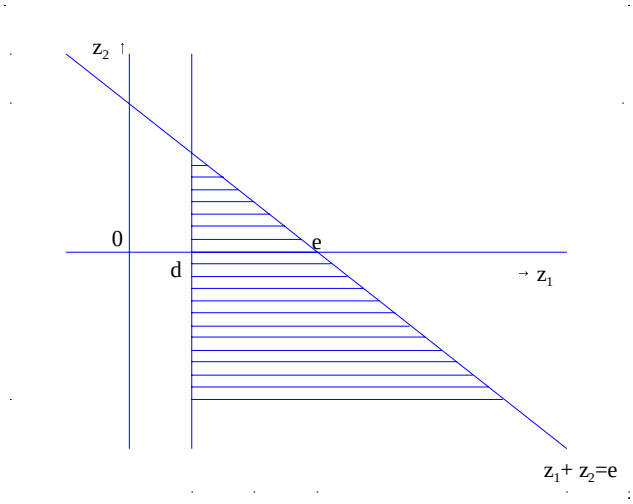
Since the combined sample leads to rejecting $H_0$ for

$$\underline{\overline{y}}_3 < m - z_\alpha \sigma/\sqrt{2n_1} \Longleftrightarrow \underline{z}_1 + \underline{z}_2 < 2\frac{m - \mu}{\sigma/\sqrt{n_1}} - z_\alpha\sqrt{2} = e$$

the power $\beta^*$ of this two-step test equals

$$\begin{aligned} \beta^*(\mu) &= P_\mu \text{ (reject } H_0) \\ &= P_\mu(\underline{z}_1 < d) + P_\mu(\underline{z}_1 \geq d \cap \underline{z}_2 < e - \underline{z}_1) \end{aligned}$$

The figure shows the area relating to the latter probability.




Using the independence of $\underline{z}_1$ and $\underline{z}_2$ then gives

$$\beta^*(\mu) = \Phi(d) + \int_d^\infty \varphi(z_1)\Phi(e - z_1)dz_1 \tag{3.8}$$

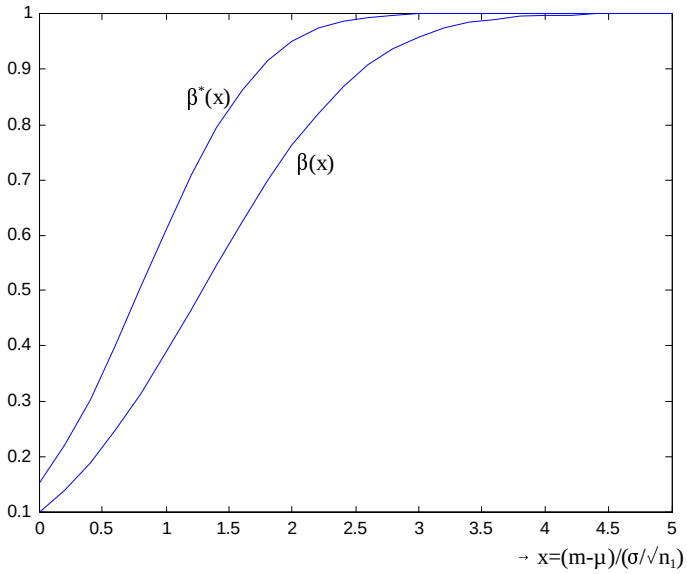
Since $\Phi(d)$ represents the power $\beta$ of the usual, one-step test procedure, the power is increased uniformly, as is intuitively clear.

In particular, the size $\alpha^*$ of the two-step test is

$$\alpha^* = \Phi(-z_\alpha) + \int_{-z_\alpha}^\infty \varphi(z)\Phi(-z_\alpha\sqrt{2} - z)dz$$

For $\alpha = 0.1$ e.g., $\alpha^* = 0.1526$. Figure 3.6 shows the two powers $\beta$ and $\beta^*$ as function of $\sqrt{n_1}(m - \mu)/\sigma$; note the large differences.

**Figure 3.6**. Power functions of one- and two-step procedures.




# 4 Extension based on sample variance

In this section, doubling of the original sample size $n_1$ depends on event $C_3 = \{\underline{s}_1^2 > c\}$. In this case, the independence of $\overline{\underline{y}}_1$ (and $\overline{\underline{y}}_2$) of $\underline{s}_1^2$ guarantees the unbiasedness of the final estimator $\overline{\underline{y}}$ for $\mu$. As in the derivation of (3.3), it follows

$$E\left[\left(\frac{\overline{\underline{y}}-\mu}{\sigma/\sqrt{n_1}}\right)^2\right] \;=\; 1+E[(\underline{z}_1+\underline{z}_2)^2/4-\underline{z}_1^2|C_3]P(C_3) = 1-\tfrac{1}{2}P(\underline{s}_1^2>c)$$

Introduce

$$b=(n_1-1)c/\sigma^2, \quad \underline{x}=(n_1-1)\underline{s}_1^2/\sigma^2 \sim \chi^2_{n_1-1}$$

and let $g_\nu$ and $G_\nu$ denote density and distribution function of $\chi_\nu$, respectively, so that

$$g_\nu(x)=\frac{1}{2^{\nu/2}\Gamma(\nu/2)}x^{\nu/2-1}e^{-x/2}, \quad x>0$$

With $\nu = n_1 - 1$ the variance of $\overline{\underline{y}}$ can be written as

$$\mathrm{Var}(\overline{\underline{y}}) = \frac{1}{2}\sigma^2[1+G_\nu(b)]/n_1 \tag{4.1}$$

Note that this formula can be derived as well from

$$\mathrm{Var}(\overline{\underline{y}}) = E[\mathrm{Var}(\overline{\underline{y}}|\underline{n})] = \sigma^2 E(1/\underline{n})$$

So, Var($\underline{\overline{y}}$) is increasing in $b$, from the limit $\sigma^2/(2n_1)$ for $b \to 0$, to $\sigma^2/n_1$ for $b \to \infty$.

As in the previous section

$$\begin{aligned}(2n_1-1)[E(\underline{s}^2)-\sigma^2] &= E\left[(n_1-1)\underline{s}_2^2 - n_1\underline{s}_1^2 + \frac{n_1}{2}(\underline{\overline{y}}_1 - \underline{\overline{y}}_2)^2 | C_3\right] P(C_3) \\ &= n_1\sigma^2 P(C_3) - \frac{n_1}{n_1-1}\sigma^2 \int_b^\infty x g_\nu(x)dx \\ &= \frac{n_1\sigma^2}{\nu}\int_b^\infty (\nu - x) g_\nu(x)dx = \frac{n_1\sigma^2}{\nu}[2x g_\nu(x)]_b^\infty\end{aligned}$$

leading to

$$E(\underline{s}^2) = \sigma^2 - \frac{2n_1}{2n_1-1} c g_\nu(b) \tag{4.2}$$

That $\underline{s}^2$ shows a negative bias is intuitively clear; it disappears for $c \to 0$ or $\infty$.

The expectation of the variance estimator can be found similarly:

$$\begin{aligned}E[\underline{\mathrm{var}}(\underline{\overline{y}})] - \frac{\sigma^2}{n_1} &= \frac{1}{2n_1} E\left[\underline{s}_3^2 - 2\underline{s}_1^2 | C_3\right] P(C_3) \\ &= \frac{1}{2n_1(2n_1-1)}\left[n_1\sigma^2 - (3n_1-1)E(\underline{s}_1^2|C_3)\right] P(C_3) \\ &= \frac{\sigma^2}{2n_1(2n_1-1)}\int_b^\infty \left[\frac{3n_1-1}{n_1-1}(\nu - x) - (2n_1-1)\right] g_\nu(x)dx \\ &= \frac{\sigma^2}{2n_1(2n_1-1)}\left[-\frac{3n_1-1}{n_1-1} 2b g_\nu(b) - (2n_1-1)(1-G_\nu(b))\right]\end{aligned}$$

Simplification gives

$$E\left[\underline{\mathrm{var}}(\underline{\overline{y}})\right] = \frac{\sigma^2}{2n_1}[1 + G_\nu(b)] - \frac{3n_1-1}{n_1(2n_1-1)} c g_\nu(b) \tag{4.3}$$

Comparison with (4.1) shows that the last term represents the bias of the variance estimator. Comparison with (4.2) shows that $B[(\underline{\mathrm{var}}\ (\underline{\overline{y}})]/B(\underline{s}^2) \approx 1.5/n_1$; note the dramatic difference with Section 3: this biasratio is now inversely proportional to the sample size.

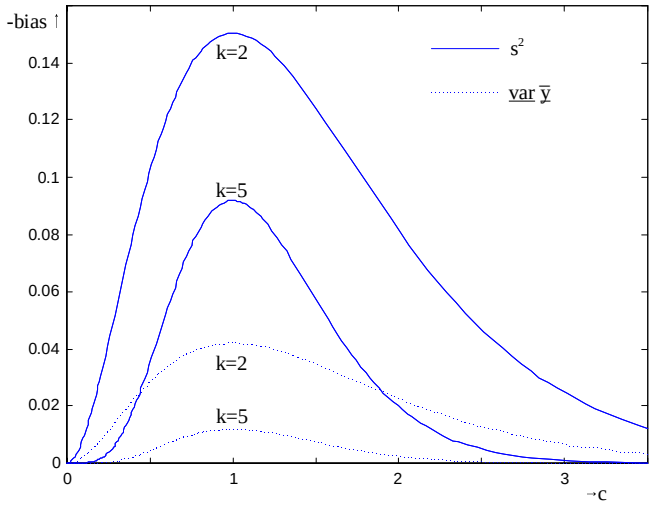
For $\sigma^2 = 1$ and $n_1 = 2k+1$, the biases can be written as

$$B(\underline{s}^2) = -\frac{2k+1}{4k+1} \cdot \frac{(kc)^k}{k!} e^{-kc}$$

$$B[\underline{\text{var}}(\overline{\underline{y}})] = -\frac{3k+1}{(2k+1)(4k+1)} \cdot \frac{(kc)^k}{k!} e^{-kc}$$

They are shown in Figure 4.1 as function of $c$.

**Figure 4.1.** Biases of $\underline{s}^2$ and $\underline{\text{var}}(\overline{\underline{y}})$ using $C_3$.




The maximum is reached for $c = 1$ throughout - not surprisingly, since here $\sigma^2 = 1$ is assumed.

It is useful as well to look at the maximum (absolute) bias. The maximum of $bg_\nu(b)$ is reached for $b = \nu$, so that

$$\max_b [cg_\nu(b)] = \sigma^2 g_\nu(\nu) = \frac{\sigma^2}{2} \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2+1)} e^{-\nu/2}$$

Stirling's formula gives for $\nu \to \infty$

$$\max_b [cg_\nu(b)] \to \frac{\sigma^2}{2\sqrt{\pi\nu}}$$

Hence, for large $n_1$, the extreme relative biases are

$$\text{ERB } (\underline{s}^2) \approx -\frac{0.282}{\sqrt{n_1}} \qquad \text{ERB } [\underline{\text{var}}\ (\overline{\underline{y}})] \approx -\frac{0.423}{\sqrt{n_1}} \tag{4.4}$$

So, both are decreasing in $n_1$.

Since the expected sample size is

$$E(\underline{n}) = n_1[2 - G_\nu(b)]$$

the loss in accuracy, compared with a fixed sample of equal expected size is now given by

$$\frac{\mathrm{Var}(\overline{\underline{y}})}{\mathrm{Var}^*(\overline{\underline{y}})} - 1 = G_\nu(b)[1 - G_\nu(b)]/2$$

So the maximum loss is 12.5%, as in Section 2.

Quite similar to the previous section, $N$ final samples were generated for $C_3$. The empirical biases are again confronted with their theoretical values in Appendix B. The agreement is quite good, although the number of final samples was chosen according to $n_1 * N = 200,000$ (or approximately). Figures 4.2 and 4.3 show histograms of the observed values of $\mathrm{var}(\overline{\underline{y}})$. Evidently, the maximum value of $\mathrm{var}(\overline{\underline{y}})$ when $C_3$ is false equals $c/n_1$.

**Figure 4.2**. Empirical distribution of $\mathrm{var}(\overline{\underline{y}})$ using $C_3$ ($n_1 = 9$; $N$ =22,000).

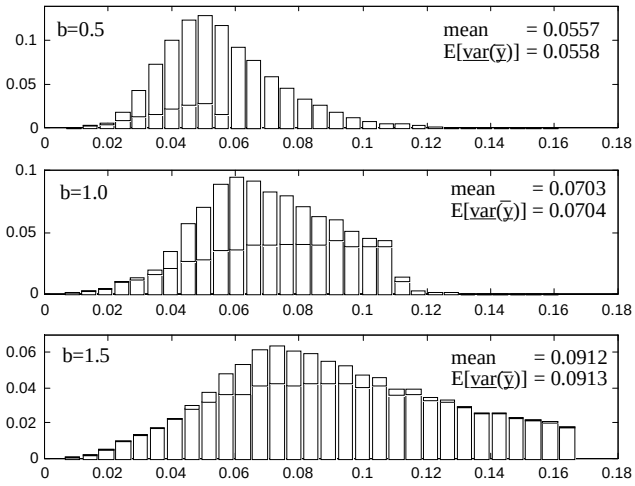
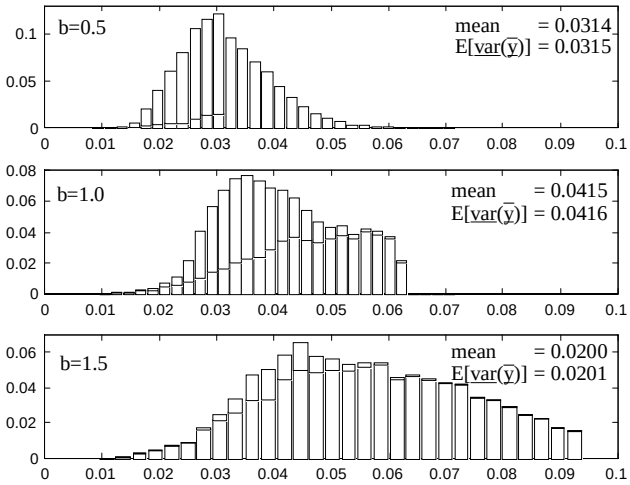


**Figure 4.3**. Empirical distribution of var($\underline{\overline{y}}$) using $C_3$ ($n_1 = 16$; $N$ =12,500).




Application of criterion $C_3$ will occur when $\mu$ has to be estimated with a given accuracy, while $\sigma^2$ is unknown. Two famous textbooks on sampling suggest to draw a pilot sample and extend this on the basis of the observed sample variance. KISH (1965, p.52) states: 'First, collect a basic sample of reasonable minimum size that might meet the demands. Then compute the results and, if the demands are not met, collect a supplementary sample of desired size. This two-step procedure can be used to obtain either a desired variance or sample size.' No mention is made of the problems connected with this two-step sequential approach. COCHRAN (1977, Section 4.7) presents more details: sample size formulae are given for different cases, the biasedness of the sample mean is noted.

# 5 Extension based on two sample variances

This section considers the case where in the first step a sample of size $n_1$ is drawn from each of two normal populations $N(\mu_i, \sigma_i^2)$, $\quad (i = 1, 2)$, independently. Extension of the samples depends on the ratio of the sample variances observed in the first step: the sample with the higher variance is increased proportionally. Hence, in the sequel stochastic step 2 sample sizes $\underline{n_2}$ will occur.

Denoting the first step variances by $\underline{s_1^2}$ and $\underline{t_1^2}$ for population 1 and 2, respectively, the extension criterion for sample 1 reads

$$C_4 = \{\underline{s_1^2} > \underline{t_1^2}\}$$

and the second step sample size $\underline{n}_2$ for population 1 becomes

$$\underline{n}_2 = \begin{cases} 0 & \text{if } C_4' \\ n_1(\underline{s}_1^2/\underline{t}_1^2 - 1) & \text{if } C_4 \end{cases}$$

Note that the last expression has to be an integer: resulting rounding errors will be neglected throughout the paper. The final sample size for population 1 can be written as

$$\underline{n} = n_1 \underline{s}_1^2 / \min(\underline{s}_1^2, \underline{t}_1^2)$$

For population 2 the sample is increased if $C_4'$ occurs, the final sample size being $n_1 \underline{t}_1^2 / \min(\underline{s}_1^2, \underline{t}_1^2)$.

This two-step procedure will be a straightforward approach when two normal sample means have to be estimated with equal accuracy: note that for $\underline{s}_1^2 = \sigma_1^2$ and $\underline{t}_1^2 = \sigma_2^2$, both final sample means have variance $\min(\sigma_i^2)/n_1$. In regression, equal variances are desirable to obtain homoskedastic models.

The final sample 1 mean $\overline{\underline{y}}$ is again unbiased for $\mu_1$, with variance

$$\mathrm{Var}(\overline{\underline{y}}) = \sigma_1^2 E(1/\underline{n})$$

Introduce $\tau = \sigma_1^2/\sigma_2^2$ and $\underline{x} = \tau \underline{t}_1^2/\underline{s}_1^2 \sim F_{\nu,\nu}$; let $f_\nu$ and $F_\nu$ denote density and distribution function of the $F_{\nu,\nu}$-distribution, so that

$$f_\nu(x) = \frac{\Gamma(\nu)}{[\Gamma(\nu/2)]^2} \frac{x^{\nu/2-1}}{(1+x)^\nu}, \quad x > 0$$

Then $C_4 = \{\underline{x} < \tau\}$ and

$$E(1/\underline{n}) = \frac{1}{n_1} P(C_4') + \frac{1}{n_1 \tau} E(\underline{x}|C_4) P(C_4)$$

so that

$$n_1 \,\mathrm{Var}(\overline{\underline{y}})/\sigma_1^2 = n_1 E(1/\underline{n}) = 1 - \int_0^\tau (1 - x/\tau) f_\nu(x) dx \tag{5.1}$$

Denote the final sample size (mean) from population 2 by $\underline{m}$ and $\overline{\underline{w}}$, respectively. Then, it follows similarly

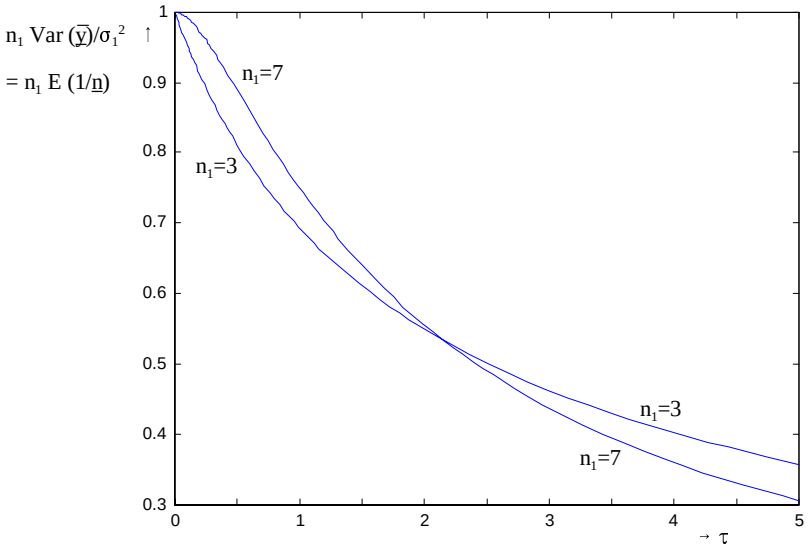
$$n_1 \,\mathrm{Var}(\overline{\underline{w}})/\sigma_2^2 = n_1 E(1/\underline{m}) = 1 - \int_0^{1/\tau} (1 - \tau x) f_\nu(x) dx$$

This expression can be found from (5.1) by substituting $1/\tau$ for $\tau$, thanks to the property $\underline{s}_1^2/(\tau \underline{t}_1^2) = 1/\underline{x} \sim F_{\nu,\nu}$. For odd $n_1$, the integral can be solved simply; some values of (5.1) are shown in the following table (first column) and figure.

**Table 5.1.** $\mathrm{Var}(\overline{\underline{y}})$ as function of $\tau$, using $C_4$.

| $n_1$ | $n_1 \mathrm{Var}(\overline{\underline{y}})/\sigma_1^2$ | $\mathrm{Var}(\overline{\underline{y}})/\mathrm{Var}^*(\overline{\underline{y}}) - 1$ |
|---|---|---|
| 3 | $\frac{1}{\tau}\ln(\tau+1)$ | – |
| 5 | $1 - \frac{\tau^2}{(\tau+1)^2}$ | $\frac{3\tau^2}{(\tau+1)^2} - \frac{\tau^2}{(\tau+1)^4}$ |
| 7 | $1 - \frac{\tau^3(2\tau+5)}{2(\tau+1)^4}$ | $\frac{5\tau^3(\tau+4)}{4(\tau+1)^4} - \frac{\tau^3(2\tau+5)(5\tau+2)}{4(\tau+1)^8}$ |
| 9 | $1 - \frac{\tau^4(3\tau^2+14\tau+21)}{3(\tau+1)^6}$ | $\frac{7\tau^4(\tau^2+6\tau+15)}{9(\tau+1)^6} - \frac{\tau^4(3\tau^2+14\tau+21)(21\tau^2+14\tau+3)}{9(\tau+1)^{12}}$ |

**Figure 5.1.** $\mathrm{Var}(\overline{\underline{y}})$ as function of $\tau$, using $C_4$.

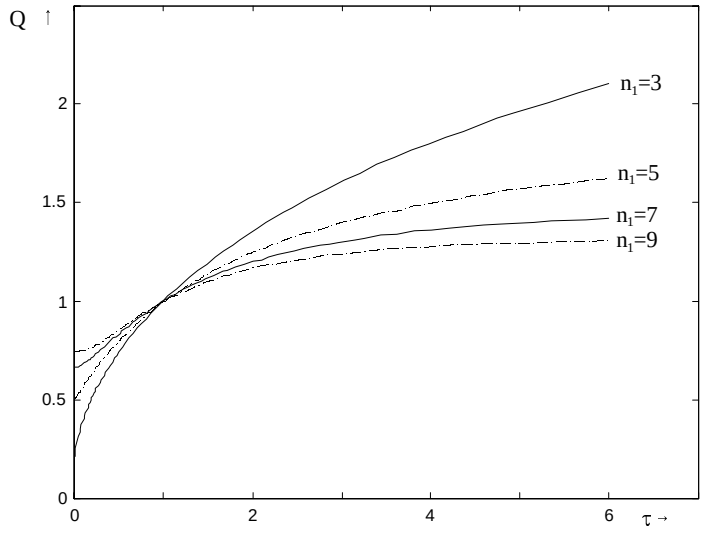



Since the goal in this section was to obtain approximately equal variances, the variance ratio

$$Q = \mathrm{Var}(\overline{\underline{y}})/\mathrm{Var}(\overline{\underline{w}}) = \tau E(1/\underline{n})/E(1/\underline{m})$$

is of great interest. Figure 5.2 shows the behavior of $Q$ for $n_1 = 3(2)9$.

**Figure 5.2.** $Q = \mathrm{Var}(\overline{\underline{y}})/\,\mathrm{Var}(\overline{\underline{w}})$ using $C_4$.




Using $q = (n_1 - 1)/2$, we conjecture that the inequalities

$$(q-1)/q < Q < q/(q-1)$$

hold generally.

The loss of accuracy due to stochastic sample size is given now by

$$\mathrm{Var}(\overline{\underline{y}})/\,\mathrm{Var}^*(\overline{\underline{y}}) - 1 = E(\underline{n})E(1/\underline{n}) - 1 \tag{5.2}$$

where the expected sample size can be found quite similarly as (5.1), using $E(\underline{x}) = \nu/(\nu-2)$ for $\nu > 2$:

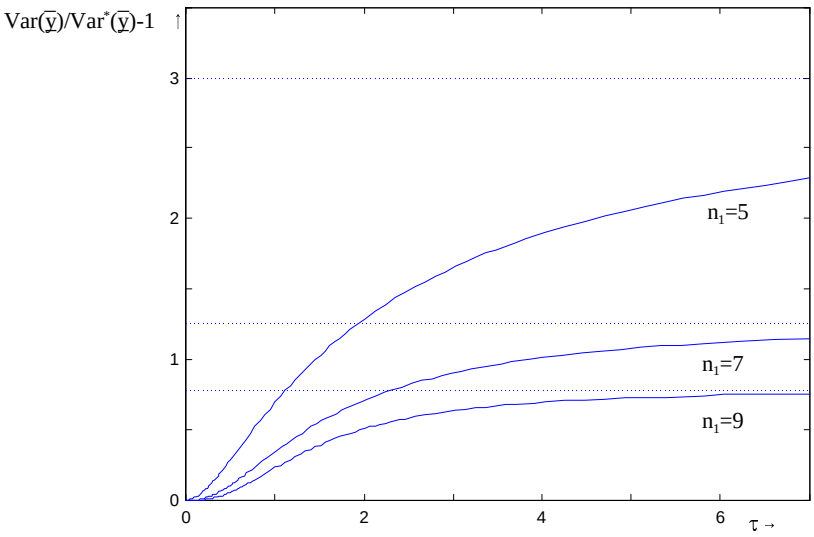
$$\frac{1}{n_1}E(\underline{n}) = \tau\frac{\nu}{\nu-2} + \int_0^{1/\tau}(1-\tau x)f_\nu(x)dx \tag{5.3}$$

In combination with the formula below (5.1) this gives the simple expression

$$\frac{1}{n_1}E(\underline{n}) + n_1E(1/\underline{m}) = 1 + \tau\frac{\nu}{\nu-2} \tag{5.4}$$

Some values of (5.2) are given in Table 5.1 as well. We conjecture that the (maximum) loss of accuracy tends to $4(\nu-1)/(\nu-2)^2$ for $\tau \to \infty$. Compare Figure 5.3. The dotted lines indicate the limiting behaviour for $\tau \to \infty$.

**Figure 5.3**. Accuracy loss from two-step procedure, using $C_4$.




Again, for the second population counterparts of (5.2)-(5.4) are obtained by substituting $1/\tau$ for $\tau$.

Expressions for the sample variances are obtained as in Section 3:

$$E(\underline{s}^2) = \sigma_1^2 + E(\underline{s}_3^2 - \underline{s}_1^2|C_4)P(C_4)$$

$$(\underline{n}-1)\underline{s}_3^2 = (n_1-1)\underline{s}_1^2 + (\underline{n}_2-1)\underline{s}_2^2 + \frac{n_1\underline{n}_2}{\underline{n}}(\overline{\underline{y}}_1 - \overline{\underline{y}}_2)^2 \tag{5.5}$$

For any pair of given step 1 sample variances $(s_1^2, t_1^2)$, $n_2$ and $n$ are fixed as well; consequently, the conditional expectation under $C_4$ of the last two terms in (5.5) totals $n_2\sigma_1^2$. Hence

$$\begin{aligned} E(\underline{s}^2) - \sigma_1^2 &= E\left[\frac{\underline{n}_2}{\underline{n}-1}(\sigma_1^2 - \underline{s}_1^2)|C_4\right]P(C_4) \\ &= E\left[\frac{\underline{n}}{\underline{n}-1}\{\sigma_1^2(1-\underline{t}_1^2/\underline{s}_1^2) - (\underline{s}_1^2 - \underline{t}_1^2)\}|C_4\right]P(C_4) \\ &\approx E\left[\sigma_1^2(1-\underline{t}_1^2/\underline{s}_1^2) - (\underline{s}_1^2 - \underline{t}_1^2)|C_4\right]P(C_4) \end{aligned}$$

by neglecting the factor $\underline{n}/(\underline{n}-1)$. Note that a slightly better approximation might be obtained by adding the factor $E(\underline{n}|C_4)/[E(\underline{n}|C_4)-1]$. As before, the first term can be rewritten as

$$\sigma_1^2 E(1-\underline{t}_1^2/\underline{s}_1^2|C_4)P(C_4) = \sigma_1^2\int_0^\tau (1-x/\tau)f_\nu(x)dx$$

To evaluate the second term, the notations $\underline{u} = (n_1 - 1)\underline{s}_1^2/\sigma_1^2 \sim \mathcal{X}_\nu^2$ and $\underline{v} = (n_1 - 1)\underline{t}_1^2/\sigma_2^2 \sim \mathcal{X}_\nu^2$ will be used, leading to

$$E(\underline{s}_1^2 - \underline{t}_1^2|C_4)P(C_4) = \frac{\sigma_1^2}{\nu}E(\underline{u} - \underline{v}/\tau|\underline{u} > \underline{v}/\tau)P(\underline{u} > \underline{v}/\tau)$$

$$= \frac{\sigma_1^2}{\nu}\int_0^\infty \int_0^{\tau u}(u - v/\tau)g_\nu(u)g_\nu(v)dvdu$$

Since

$$\int_0^{\tau u}(u - v/\tau)g_\nu(v)dv = \frac{1}{\tau}\int_0^{\tau u}[\tau u - \nu + (\nu - v)]g_\nu(v)dv$$

$$= \frac{1}{\tau}(\tau u - \nu)G_\nu(\tau u) + 2ug_\nu(\tau u)$$

this leads to

$$\frac{\nu}{\sigma_1^2}E(\underline{s}_1^2 - \underline{t}_1^2|C_4)P(C_4) =$$

$$\int_0^\infty (u - \nu/\tau)g_\nu(u)G_\nu(\tau u)du + 2\int_0^\infty ug_\nu(u)g_\nu(\tau u)du \tag{5.6}$$

For simplicity, only odd $n_1 = 2k + 1$ will be considered again. Using the property

$$G_{2k}(x) = 1 - 2\sum_{j=1}^{k} g_{2j}(x)$$

the right-hand side of (5.6) can be written as

$$\int_0^\infty (u - \nu/\tau)g_{2k}(u)du + \frac{2\nu}{\tau}\int_0^\infty g_2(\tau u)g_{2k}(u)du +$$

$$+2\sum_{j=1}^{k-1}\int_0^\infty [\nu g_{2(j+1)}(\tau u)/\tau - ug_{2j}(\tau u)]g_{2k}(u)du$$

Further simplification leads to

$$\nu\frac{\tau - 1}{\tau} + \frac{\nu}{\tau(\tau + 1)^k} + 2\sum_{j=1}^{k-1}(k - j)\begin{pmatrix} k + j - 1 \\ k - 1 \end{pmatrix}\frac{\tau^{j-1}}{(\tau + 1)^{k+j}}$$

resulting for $\nu = 2k$ in

$$E(\underline{s}_1^2 - \underline{t}_1^2|C_4)P(C_4)/\sigma_1^2 = \frac{\tau - 1}{\tau} + \sum_{j=0}^{k-1}\frac{k - j}{k}\begin{pmatrix} k + j - 1 \\ k - 1 \end{pmatrix}\frac{\tau^{j-1}}{(\tau + 1)^{k+j}} \tag{5.7}$$

Note that symmetry properties lead from (5.7) to

$$E(\underline{t}_1^2 - \underline{s}_1^2|C_4')P(C_4')/\sigma_1^2 = \frac{1-\tau}{\tau} + \sum_{j=0}^{k-1} \frac{k-j}{k} \begin{pmatrix} k+j-1 \\ k-1 \end{pmatrix} \frac{\tau^k}{(\tau+1)^{k+j}}$$

Subtracting these two gives

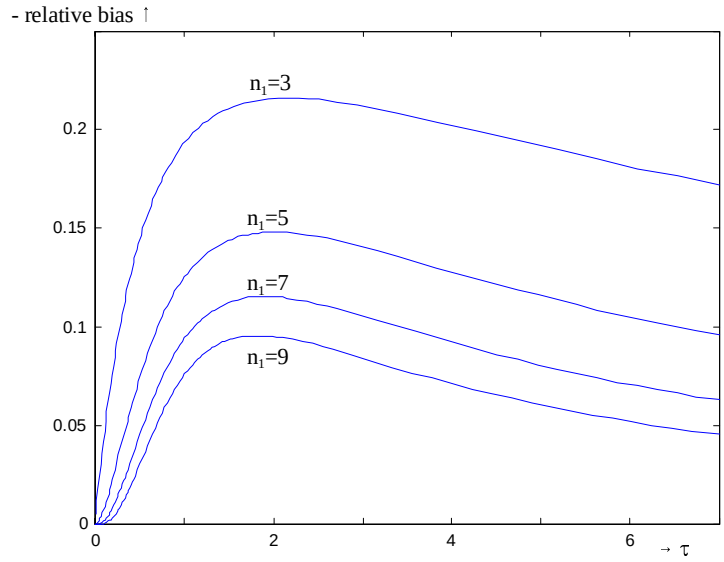
$$\sum_{j=0}^{k-1} \frac{k-j}{k} \begin{pmatrix} k+j-1 \\ k-1 \end{pmatrix} \frac{\tau^k - \tau^{j-1}}{(\tau+1)^{k+j}} = E(\underline{s}_1^2 - \underline{t}_1^2)/\sigma_1^2 = \frac{\tau-1}{\tau} \tag{5.8}$$

As a check, Appendix A gives a direct proof of the equality of the outer terms in (5.8). Table 5.2 and Figure 5.4 show some detailed results.

**Table 5.2** Relative bias of $\underline{s}^2$ using $C_4$.

| $n_1$ | $E(\underline{s}_1^2 - \underline{t}_1^2\|C_4)P(C_4)/\sigma_1^2$ | $E(\underline{s}^2)/\sigma_1^2 - 1 \approx$ |
|---|---|---|
| 3 | $\frac{\tau}{\tau+1}$ | $\frac{1}{\tau+1} - \frac{1}{\tau}\ln(\tau+1)$ |
| 5 | $\frac{\tau^2(\tau+2)}{(\tau+1)^3}$ | $-\frac{\tau^2}{(\tau+1)^3}$ |
| 7 | $\frac{\tau^3(\tau^2+4\tau+5)}{(\tau+1)^5}$ | $-\frac{\tau^3(\tau+5)}{2(\tau+1)^5}$ |
| 9 | $\frac{\tau^4(\tau^3+6\tau^2+14\tau+14)}{(\tau+1)^7}$ | $-\frac{\tau^4(\tau^2+7\tau+21)}{3(\tau+1)^7}$ |

**Figure 5.4**. Relative bias of $\underline{s}^2$ using $C_4$.




For $\tau \to 0$ or $\tau \to \infty$, the bias of $\underline{s}^2$ disappears, but even for reasonably high values of $\tau$, the bias is considerable. Throughout, the bias is negative; this is intuitively clear, since additional observations are taken only if $\underline{s}_1^2$ exceeds $\underline{t}_1^2$. High values of $\underline{s}_1^2$, obtained by random phenomena then are corrected downwards.

For evaluating the variance estimator $\underline{\text{var}}(\overline{\underline{y}})$, starting point is

$$E[\underline{\text{var}}(\overline{\underline{y}})] = E(\underline{s}_1^2/n_1) + E(\underline{s}_3^2/\underline{n} - \underline{s}_1^2/n_1|C_4)P(C_4)$$

Using (5.5) and the argument just below this formula gives

$$E[\underline{\text{var}}(\overline{\underline{y}})] - \sigma_1^2/n_1 = E\left[\frac{1}{\underline{n}(\underline{n}-1)}\left\{\underline{n}_2\sigma_1^2 + (n_1-1)\underline{s}_1^2\right\} - \underline{s}_1^2/n_1|C_4\right]P(C_4)$$

Replacing $\underline{n}-1$ in the denominator by $\underline{n}$ once more, and re-using the variables $\underline{x}$, $\underline{u}$ and $\underline{v}$ leads to

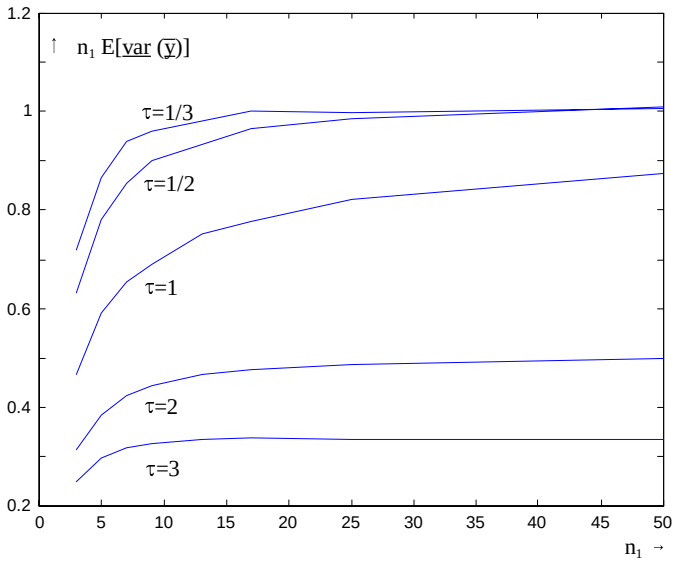
$$\frac{n_1}{\sigma_1^2}E[\underline{\text{var}}(\overline{\underline{y}})] - 1 = E[\underline{x}/\tau - \underline{x}^2/\tau^2|C_4]P(C_4) - E\left[\frac{\underline{u}}{n_1-1} - \frac{\underline{v}^2}{n_1\tau\underline{u}}|C_4\right]P(C_4)$$

In principle, this expression can be simplified in the same vein as earlier $E(\underline{s}^2)$. We plan to do so in the near future; for the moment, we satisfy ourselves with the simulation results.

$N$ pairs of final samples from the normal distributions $N(0,1)$ and $N(0,1/\tau)$ were generated with sizes $\underline{n}$ and $\underline{m}$ determined by the $C_4$ criterion; for each pair of final samples the values of $\overline{y}$, $s^2$, var($\underline{\overline{y}}$) and $\overline{w}$, $t^2$, var($\underline{\overline{w}}$) were calculated and stored separately for $C_4$ false/true. Furthermore, the intermediate step 1 values were stored in these files too. All combinations of $\tau \in \{1,2,3\}$ and $n_1 \in \{3,5,7,9,13,17,25,50\}$ were used. The number $N$ of pairs of final samples depends on $n_1$ as before; $n_1 * N$ varies between 72,000 and 150,000. In the simulation program an upperbound for $\underline{n}$ and $\underline{m}$ of 1000 was used. In particular with $n_1 = 3$ this upperbound was reached frequently (323 times for $\tau = 1$, 385 for $\tau = 2$, 469 for $\tau = 3$; without upperbounding the respective maximum values of $\underline{n}$ or $\underline{m}$ would be 120085, 734341 and 673274 for these cases). For the larger values of $n_1$ this upperbounding can be neclected for all practical purposes.

The variance of the statistic $\overline{y}$ and the average values of the statistics $s^2$ and var($\underline{\overline{y}}$) were calculated as well as the relative biases of $s^2$ and var($\underline{\overline{y}}$). The corresponding values calculated from the statistics $\overline{w}$, $t^2$, and var($\underline{\overline{w}}$) for the cases $\tau = 2$ and $\tau = 3$ were used to obtain similar results for the cases $\tau = 1/2$ and $\tau = 1/3$. Only some summarizing figures are presented here; see Appendix C for detailed tables. Figure 5.5 shows the expectation of the variance estimator.

**Figure 5.5.** Empirical values of $n_1 E[\underline{\text{var}}(\underline{\overline{y}})]$ as function of $n_1$, using $C_4$.




The picture confirms that for large $n_1$, Var($\underline{\overline{y}}$) approximates $\sigma_1^2/n_1$ for small $\tau$ and $\sigma_1^2/\tau$ for large $\tau$.

The relative biases of $\underline{s}^2$ and $\underline{\text{var}}(\underline{\overline{y}})$ as a function of $n_1$ are plotted in Figures 5.6 and 5.7, respectively. Note that the curve for $\tau = 2$ in Figure 5.6 lies above the others; that agrees with our finding in Figure 5.4 that the extreme relative bias of $\underline{s}^2$ occurs for $\tau \approx 2$.

**Figure 5.6.** Empirical relative bias of $\underline{s}^2$ as function of $n_1$, using $C_4$.

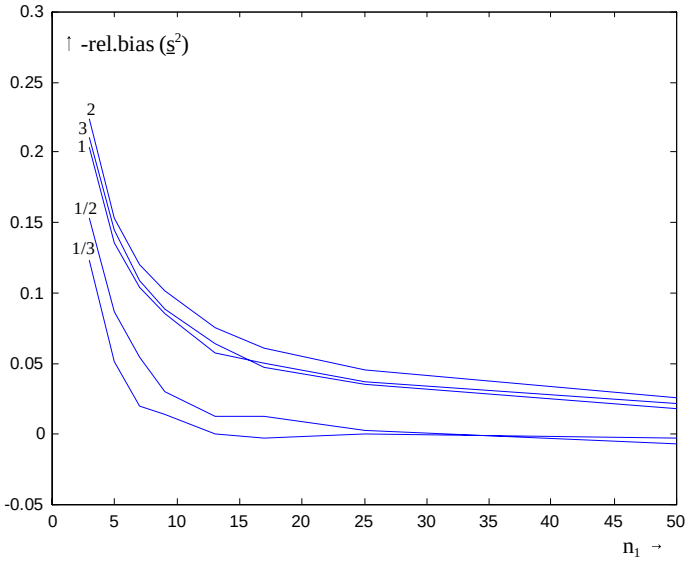
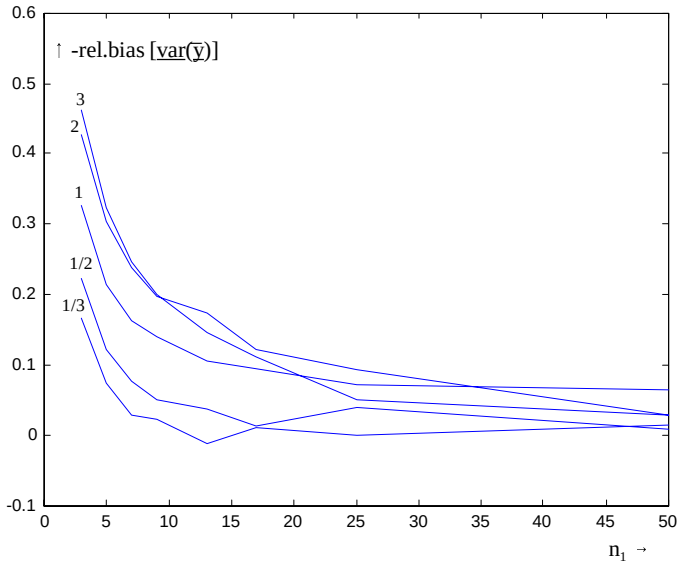



**Figure 5.7.** Empirical relative bias of $\underline{\text{var}}(\underline{\overline{y}})$ as function of $n_1$, using $C_4$.



The simulated values agree with the theoretical ones for small $n_1$. The biases are sizable; even for $n_1$ as large as 50, the variance estimator may have a bias of about 6%.

In KLEIJNEN & VAN GROENENDAAL (1995) the more general situation of $n$ normal populations was considered, in stead of just two. (We have restricted ourselves to those features of their paper that relate to ours; for convenience, their notation is adapted to ours.) From each population an independent step 1 sample of size $n_1$ is drawn. Use the subscript $i$ to indicate each population, so that $\underline{s}_{1i}^2$, for example, denotes the variance of the first $n_1$ observations in population $i$. Then the final sample sizes $\underline{n}_i$ are determined according to

$$\underline{n}_i = n_1 \underline{s}_{1i}^2 / \min_i \underline{s}_{1i}^2$$

for $i = 1, 2, ..., n$. The standard variance estimators

$$\underline{\mathrm{var}}(\underline{\overline{y}}_i) = \underline{s}_{1i}^2 / \underline{n}_i$$

are used again. It is stated (KLEIJNEN & VAN GROENENDAAL (1995), p.92) that these variance estimators are unbiased for $\mathrm{Var}(\underline{\overline{y}}_i)$. It will be clear from our analysis that this statement is incorrect. (The first present author has to admit that - at that time - he was convinced of the unbiasedness of $\underline{\mathrm{var}}(\underline{\overline{y}}_i)$. Let us consider that as an illustration of the progress of scientific insights.)

# 6 Discussion

In this paper the effects of extension of a random sample of fixed size were studied in some detail. Although only simple situations were considered, the results are surprising and even disturbing. For example, it was quite a shock for us that the final sample variance is biased when extension criterion $C_4$ is used. Some results for the less complex cases $C_1 - C_3$ are summarized in Table 6.1; to highlight the *possible* effects of the use of standard methods, worst-case scenarios are presented.

**Table 6.1.** Maximum loss of accuracy and extreme relative biases (ERB).

| Crit. | Max.increase $\mathrm{Var}(\underline{\overline{y}})$ | ERB($\underline{\overline{y}}$) | ERB($\underline{s}^2$) | ERB[$\underline{\mathrm{var}}(\underline{\overline{y}})$] |
|---|---|---|---|---|
| $C_1$ | 12.5% | 0 | 0 | 0 |
| $C_2$ | 40.2% | $-\frac{0.20\sigma/\mu}{\sqrt{n_1}}$ | $\frac{0.06}{n_1}$ | 0.18 |
| $C_3$ | 12.5% | 0 | $-\frac{0.28}{\sqrt{n_1}}$ | $-\frac{0.42}{\sqrt{n_1}}$ |

For criterion $C_4$, the step two sample size becomes stochastic, leading to a much more complicated analysis. Therefore, numerical results are summarized in Table 6.2 for the most important statistic, the final variance estimator.

**Table 6.2.** Relative bias (in %) of $\underline{\mathrm{var}}(\underline{\overline{y}})$ using $C_4$.

| $n_1$ | 5 | 9 | 13 | 17 | 25 | 50 |
|---|---|---|---|---|---|---|
| $\tau$ | | | | | | |
| $\frac{1}{3}$ | −7.52 | −2.29 | 1.20 | −1.01 | 0.00 | −1.47 |
| $\frac{1}{2}$ | −12.26 | −5.03 | −3.70 | −1.39 | −3.90 | −0.98 |
| 1 | −21.47 | −13.92 | −10.54 | −9.50 | −7.32 | −6.42 |
| 2 | −30.51 | −19.84 | −17.24 | −12.23 | −9.30 | −2.91 |
| 3 | −32.34 | −19.91 | −14.57 | −11.16 | −4.96 | −2.90 |

Due to the relatively small number of replications for large $n_1$, the figures in the latter columns are not too reliable.

Throughout this paper an unconditional approach was followed, in the sense that the complete two-step procedure depicted in Figure 1.1 was evaluated beforehand. At the end of step 1, however, the final sample size is a known number $n$; using this additional information leads to the so-called conditional approach, where the final sample is evaluated under the condition $\{\underline{n} = n\}$. Since the variance estimator $\underline{\mathrm{var}}(\underline{\overline{y}})$ in particular heavily depends on the final sample size, it might be conjectured that this variance estimator performs better conditionally.

Indeed, for criterion $C_1$ this conditional approach has an advantage: the loss of accuracy of course disappears. For all three remaining cases, however, this conjecture proves to be false. To show this briefly, the situation that $C_i$ is false will be considered, leading to $\underline{n}= n_1$ and enabling comparison for i=2,3 and 4.

The short Table 6.3 presents the relative bias of $\underline{\mathrm{var}}(\underline{\overline{y}})$ calculated from the simulation experiment mentioned before. Both the unconditional and the conditional situation are considered; in the latter case relative bias is defined as $E\left[\underline{\mathrm{var}}(\underline{\overline{y}}|\underline{n} = n_1)\right] / \mathrm{Var}(\underline{\overline{y}}|\underline{n} = n_1) - 1$.

**Table 6.3.** Unconditional and conditional (given $\underline{n} = n_1$) relative bias of $\underline{\mathrm{var}}(\underline{\overline{y}})$.

| $n_1$ | 9 | | 25 | |
|---|---|---|---|---|
| Criterion | uncond. | cond. | uncond. | cond. |
| $C_2(d=0)$ | 0.06 | 1.78 | 0.06 | 1.81 |
| $C_3(b=1)$ | −0.19 | −0.34 | −0.11 | −0.23 |
| $C_4(\tau=1)$ | −0.15 | −0.29 | −0.07 | −0.17 |

Throughout, the conditional biases are more serious.

We plan to extend this paper in a few directions. First of all, some remaining loose ends in Section 5 need attention: the theoretical evaluation of $E\left[\underline{\mathrm{var}}(\underline{\overline{y}})\right]$ and the two conjectures regarding $\mathrm{Var}(\underline{\overline{y}})$ in relation to $\mathrm{Var}(\underline{\overline{w}})$ and $\mathrm{Var}^*(\underline{\overline{y}})$, respectively. An essential assumption up to now was normality: the only distribution guaranteeing independence of mean and variance in random samples. It is to be expected therefore that the effects of sample extension will even be greater for other distrubutions. So, secondly, we will repeat our simulation experiments for gamma distributed variables.

Thirdly, extending samples occurs in more ways than the four considered here. We mention two interesting possibilities, related to criteria $C_3$ and $C_4$.

a) In Section 4, a large observed variance led to doubling of the sample size. In stead, one may wish to approximate a prescribed precision $a$ by taking the (now stochastic) final sample size $\underline{n} = \underline{s}_1^2/a$.

b) In Section 5, an alternative choice for the final sample sizes is

$\underline{n} = n_1\underline{s}_1/\min(\underline{s}_1, \underline{t}_1)$, $\underline{m} = n_1\underline{t}_1/\min(\underline{s}_1, \underline{t}_1)$

The argument is that for $\underline{s}_1^2 = \sigma_1^2$, $\underline{t}_1^2 = \sigma_2^2$ this gives $n/m = \sigma_1/\sigma_2$, which minimizes $\mathrm{Var}(\underline{\overline{y}} - \underline{\overline{w}})$ for fixed $n + m$.

Of course, in stead of using standard estimation methods, alternatives could be developed that are more appropriate in case of stochastic sample size. Hoever, the main message of this paper is to warn practitioners that extending the original sample requires delicate care. Note that many of the problems discussed here arise in a lot of similar situations. In particular stochastic final sample sizes are a recurring phenomenon We quote two sentences from SÄRNDAL et al. (1992):

- 'BE (Bernoulli) sampling is often considerably less precise than simple random sampling ... explained by the variability of the size of the BE sample.' (p.55),

- ' there is a nonnegligible loss of precision caused by the lack of control of domain sample size.' (p.397).

# Literature

COCHRAN, W.G. (1977), Sampling techniques, $3^d$ ed., Wiley.

FELLER,W. (1968), An introduction to probability theory and its applications, Volume 1, Wiley.

KISH, L. (1965), Survey sampling, Wiley.

KLEIJNEN, J.P.C. & W. van GROENENDAAL (1995), Two-stage versus sequential sample-size determination in regression analysis of simulation experiments, American Journal of Mathematical and Management Sciences 15, p.83-114.

SÄRNDAL, C.-E., B. SWENSSON & J. WRETMAN (1992), Model assisted survey sampling, Springer.

# A Direct proof of (5.8)

Equality (5.8) is equivalent to

$$\sum_{j=0}^{k-1} \frac{k-j}{k} \left( \begin{array}{c} k+j-1 \\ j \end{array} \right) (\tau+1)^{k-j-1} \frac{\tau^{k+1}-\tau^j}{\tau-1} = (\tau+1)^{2k-1}$$

Since the left-hand side can be written as

$$\sum_{j=0}^{k-1} \sum_{i=j}^{k} \sum_{l=0}^{k-j-1} \frac{k-j}{k} \left( \begin{array}{c} k+j-1 \\ j \end{array} \right) \left( \begin{array}{c} k-j-1 \\ l \end{array} \right) \tau^{i+l} =$$

$$\sum_{n=0}^{2k-1} \tau^n \sum_{l=\max(0,n-k)}^{\min(n,k-1)} \sum_{j=0}^{\min(n,k-1)-l} \frac{k-j}{j} \left( \begin{array}{c} k+j-1 \\ j \end{array} \right) \left( \begin{array}{c} k-j-1 \\ l \end{array} \right)$$

this equation is again equivalent to

$$\sum_{l=\max(0,n-k)}^{\min(n,k-1)} \sum_{j=0}^{\min(n,k-1)-l} \frac{k-j}{j} \left( \begin{array}{c} k+j-1 \\ j \end{array} \right) \left( \begin{array}{c} k-j-1 \\ l \end{array} \right) = \left( \begin{array}{c} 2k-1 \\ n \end{array} \right) \qquad \text{(A.1)}$$

for $0 \leq n \leq 2k-1$. It is easily seen that the left-hand side is identical for $n$ and $2k-1-n$; hence, it suffices to proof (A.1) for $0 \leq n \leq k-1$. Now,

$$\sum_{l=0}^{n} \sum_{j=0}^{n-l} \frac{k-j}{j} \left( \begin{array}{c} k+j-1 \\ j \end{array} \right) \left( \begin{array}{c} k-j-1 \\ l \end{array} \right)$$

$$= \sum_{l=0}^{n} \sum_{j=0}^{l} \frac{k-j}{k} \left( \begin{array}{c} k+j-1 \\ j \end{array} \right) \left( \begin{array}{c} k-j-1 \\ l-j \end{array} \right)$$

$$= \sum_{l=0}^{n} \sum_{j=0}^{l} \left( \begin{array}{c} k+j-1 \\ j \end{array} \right) \left( \begin{array}{c} k-j-1 \\ l-j \end{array} \right) - \sum_{l=0}^{n-1} \sum_{j=0}^{l} \left( \begin{array}{c} k+j \\ j \end{array} \right) \left( \begin{array}{c} k-j-2 \\ l-j \end{array} \right)$$

Use of relation (12.16) in FELLER (1968)

$$\sum_{j=0}^{l} \left( \begin{array}{c} a+j-1 \\ j \end{array} \right) \left( \begin{array}{c} b+l-j-1 \\ l-j \end{array} \right) = \left( \begin{array}{c} a+b+l-1 \\ l \end{array} \right) \qquad \text{(A.2)}$$

then gives the desired result.

# B Simulation results using $C_2$ or $C_3$.

Table B.1 shows the biases of $\underline{\overline{y}}$, $\underline{s}^2$ and $\underline{\text{var}}(\underline{\overline{y}})$, if extension criterion $C_2$ is applied. The first (bold) columns contain the theoretical values, the second columns contain the values obtained by Monte Carlo simulation; the respective values of $N$ are displayed within parentheses. Generally, the simulated values approach the theoretical values closely.

**Table B.1**. Theoretical (bold) and simulated biases using $C_2$.

| | $d$ | $E(\underline{\overline{y}}) - \mu$ | | $E(\underline{s}^2) - \sigma^2$ | | $E[\underline{\text{var}}(\underline{\overline{y}})] - \text{Var}(\underline{\overline{y}})$ | |
|---|---|---|---|---|---|---|---|
| $n_1 = 4$ | $-1$ | $\mathbf{-0.0605}$ | $-0.0588$ | $\mathbf{-0.0173}$ | $-0.0165$ | $\mathbf{-0.0475}$ | $-0.0475$ |
| (250,000) | 0 | $\mathbf{-0.0997}$ | $-0.0989$ | $\mathbf{0}$ | 0.0003 | $\mathbf{0}$ | $-0.0002$ |
| | 1 | $\mathbf{-0.0605}$ | $-0.0602$ | $\mathbf{0.0173}$ | 0.0186 | $\mathbf{0.0475}$ | 0.0478 |
| $n_1 = 9$ | $-1$ | $\mathbf{-0.0403}$ | $-0.0385$ | $\mathbf{-0.0071}$ | $-0.0059$ | $\mathbf{-0.0206}$ | $-0.0206$ |
| (110,000) | 0 | $\mathbf{-0.0665}$ | $-0.0655$ | $\mathbf{0}$ | 0.0010 | $\mathbf{0}$ | $-0.0000$ |
| | 1 | $\mathbf{-0.0403}$ | $-0.0404$ | $\mathbf{0.0071}$ | 0.0078 | $\mathbf{0.0206}$ | 0.0206 |
| $n_1 = 16$ | $-1$ | $\mathbf{-0.0302}$ | $-0.0282$ | $\mathbf{-0.0039}$ | $-0.0027$ | $\mathbf{-0.0115}$ | $-0.0115$ |
| (62,500) | 0 | $\mathbf{-0.0499}$ | $-0.0489$ | $\mathbf{0}$ | 0.0015 | $\mathbf{0}$ | 0.0000 |
| | 1 | $\mathbf{-0.0302}$ | $-0.0296$ | $\mathbf{0.0039}$ | 0.0045 | $\mathbf{0.0115}$ | 0.0116 |
| $n_1 = 25$ | $-1$ | $\mathbf{-0.0242}$ | $-0.0228$ | $\mathbf{-0.0025}$ | $-0.0012$ | $\mathbf{-0.0073}$ | $-0.0073$ |
| (40,000) | 0 | $\mathbf{-0.0399}$ | $-0.0389$ | $\mathbf{0}$ | 0.0003 | $\mathbf{0}$ | $-0.0001$ |
| | 1 | $\mathbf{-0.0242}$ | $-0.0238$ | $\mathbf{0.0025}$ | 0.0029 | $\mathbf{0.0073}$ | 0.0073 |

Similar results for extension criterion $C_3$ are presented in Table B.2. Although the number of replications is five times smaller, the agreement between the theoretical and empirical biases is quite good again.

**Table B.2**. Theoretical (bold) and simulated biases using $C_3$.

| | $b$ | $E(\underline{\overline{y}}) - \mu$ | | $E(\underline{s}^2) - \sigma^2$ | | $E[\underline{\text{var}}(\underline{\overline{y}})] - \text{Var}(\underline{\overline{y}})$ | |
|---|---|---|---|---|---|---|---|
| $n_1 = 4$ | 0.5 | **0** | 0.0004 | **−0.1319** | −0.1350 | **−0.0453** | −0.0456 |
| (50,000) | 1.0 | **0** | 0.0010 | **−0.1762** | −0.1767 | **−0.0606** | −0.0604 |
| | 1.5 | **0** | 0.0019 | **−0.1529** | −0.1546 | **−0.0526** | −0.0531 |
| $n_1 = 9$ | 0.5 | **0** | 0.0007 | **−0.0478** | −0.0505 | **−0.0077** | −0.0078 |
| (22,000) | 1.0 | **0** | 0.0021 | **−0.1034** | −0.1044 | **−0.0166** | −0.0167 |
| | 1.5 | **0** | 0.0023 | **−0.0709** | −0.0723 | **−0.0114** | −0.0116 |
| $n_1 = 16$ | 0.5 | **0** | 0.0011 | **−0.0175** | −0.0204 | **−0.0016** | −0.0017 |
| (12,500) | 1.0 | **0** | 0.0024 | **−0.0744** | −0.0772 | **−0.0068** | −0.0069 |
| | 1.5 | **0** | 0.0031 | **−0.0366** | −0.0387 | **−0.0034** | −0.0034 |
| $n_1 = 25$ | 0.5 | **0** | 0.0016 | **−0.0057** | −0.0087 | **−0.0003** | −0.0004 |
| (8,000) | 1.0 | **0** | 0.0011 | **−0.0584** | −0.0608 | **−0.0035** | −0.0034 |
| | 1.5 | **0** | 0.0024 | **−0.0188** | −0.0220 | **−0.0011** | −0.0012 |

# C Simulation results using $C_4$.

Table C.1 shows results for small $n_1$ when extension criterion $C_4$ is applied.The simulated values of $\mathrm{Var}(\overline{y})$ hardly differ from their theoretical counterparts (bold). This holds for $E(\underline{s}^2)$ as well, confirming the approximate expressions in Table 5.2. For $E[\underline{\mathrm{var}}(\overline{y})]$ no theoretical values are available; the simulated values are decreasing in $\tau$ and in $n_1$.

**Table C.1**. Theoretical (bold) and simulated results using $C_4$; small $n_1$.

| | | | | | | | relative bias | |
|---|---|---|---|---|---|---|---|---|
| | $\tau$ | $\mathrm{Var}(\overline{y})$ | | $E(\underline{s}^2)$ | | $E[\underline{\mathrm{var}}(\overline{y})]$ | $\underline{s}^2$ | $\underline{\mathrm{var}}(\overline{y})$ |
| $n_1 = 3$ | $\frac{1}{3}$ | **0.2877** | 0.2871 | **0.8870** | 0.8763 | 0.2397 | −0.1237 | −0.1668 |
| (50,000) | $\frac{1}{2}$ | **0.2703** | 0.2696 | **0.8557** | 0.8472 | 0.2100 | −0.1528 | −0.2231 |
| | 1 | **0.2310** | 0.2326 | **0.8069** | 0.7963 | 0.1553 | −0.2037 | −0.3277 |
| | 2 | **0.1831** | 0.1852 | **0.7840** | 0.7758 | 0.1049 | −0.2242 | −0.4271 |
| | 3 | **0.1540** | 0.1526 | **0.7879** | 0.7893 | 0.0828 | −0.2107 | −0.4623 |
| $n_1 = 5$ | $\frac{1}{3}$ | **0.1875** | 0.1869 | **0.9531** | 0.9483 | 0.1734 | −0.0517 | −0.0752 |
| (22,000) | $\frac{1}{2}$ | **0.1778** | 0.1780 | **0.9259** | 0.9134 | 0.1560 | −0.0866 | −0.1226 |
| | 1 | **0.1500** | 0.1507 | **0.8750** | 0.8646 | 0.1178 | −0.1354 | −0.2147 |
| | 2 | **0.1111** | 0.1128 | **0.8519** | 0.8470 | 0.0772 | −0.1530 | −0.3051 |
| | 3 | **0.0875** | 0.0866 | **0.8594** | 0.8548 | 0.0592 | −0.1452 | −0.3234 |
| $n_1 = 7$ | $\frac{1}{3}$ | **0.1381** | 0.1362 | **0.9766** | 0.9807 | 0.1341 | −0.0193 | −0.0290 |
| (12,500) | $\frac{1}{2}$ | **0.1323** | 0.1344 | **0.9547** | 0.9456 | 0.1222 | −0.0544 | −0.0763 |
| | 1 | **0.1116** | 0.1110 | **0.9063** | 0.8959 | 0.0935 | −0.1041 | −0.1622 |
| | 2 | **0.0794** | 0.0790 | **0.8848** | 0.8800 | 0.0604 | −0.1200 | −0.2393 |
| | 3 | **0.0600** | 0.0609 | **0.8945** | 0.8909 | 0.0452 | −0.1091 | −0.2467 |
| $n_1 = 9$ | $\frac{1}{3}$ | **0.1090** | 0.1095 | **0.9871** | 0.9861 | 0.1065 | −0.0139 | −0.0229 |
| (8,000) | $\frac{1}{2}$ | **0.1053** | 0.1034 | **0.9698** | 0.9700 | 0.1000 | −0.0300 | −0.0503 |
| | 1 | **0.0891** | 0.0898 | **0.9245** | 0.9147 | 0.0767 | −0.0853 | −0.1392 |
| | 2 | **0.0615** | 0.0630 | **0.9049** | 0.8984 | 0.0493 | −0.1016 | −0.1984 |
| | 3 | **0.0452** | 0.0450 | **0.9160** | 0.9113 | 0.0362 | −0.0887 | −0.1991 |

Since no theoretical expressions were derived for large $n_1$, Table C.2 only presents simulation results.

**Table C.2**. Theoretical (bold) and simulated results using $C_4$; larger $n_1$.

| | | | | | relative bias | |
|---|---|---|---|---|---|---|
| | $\tau$ | $\mathrm{Var}(\overline{y})$ | $E(\underline{s}^2)$ | $E[\underline{\mathrm{var}}(\overline{y})]$ | $\underline{s}^2$ | $\underline{\mathrm{var}}(\overline{y})$ |
| $n_1 = 13$ | $\frac{1}{3}$ | 0.0753 | 1.0002 | 0.0762 | 0.0002 | 0.0120 |
| (7,000) | $\frac{1}{2}$ | 0.0756 | 0.9878 | 0.0728 | −0.0122 | −0.0370 |
| | 1 | 0.0645 | 0.9425 | 0.0577 | −0.0575 | −0.1054 |
| | 2 | 0.0435 | 0.9247 | 0.0360 | −0.0753 | −0.1724 |
| | 3 | 0.0302 | 0.9356 | 0.0258 | −0.0644 | −0.1457 |
| $n_1 = 17$ | $\frac{1}{3}$ | 0.0594 | 1.0029 | 0.0588 | 0.0029 | −0.0101 |
| (6,000) | $\frac{1}{2}$ | 0.0576 | 0.9878 | 0.0568 | −0.0122 | −0.0139 |
| | 1 | 0.0505 | 0.9491 | 0.0457 | −0.0509 | −0.0950 |
| | 2 | 0.0319 | 0.9395 | 0.0280 | −0.0605 | −0.1223 |
| | 3 | 0.0224 | 0.9520 | 0.0199 | −0.0480 | −0.1116 |
| $n_1 = 25$ | $\frac{1}{3}$ | 0.0399 | 1.0002 | 0.0399 | 0.0002 | 0.0000 |
| (4,000) | $\frac{1}{2}$ | 0.0410 | 0.9980 | 0.0394 | −0.0020 | −0.0390 |
| | 1 | 0.0355 | 0.9629 | 0.0329 | −0.0371 | −0.0732 |
| | 2 | 0.0215 | 0.9546 | 0.0195 | −0.0454 | −0.0930 |
| | 3 | 0.0141 | 0.9650 | 0.0134 | −0.0350 | −0.0496 |
| $n_1 = 50$ | $\frac{1}{3}$ | 0.0204 | 1.0026 | 0.0201 | 0.0026 | −0.0147 |
| (2,000) | $\frac{1}{2}$ | 0.0204 | 1.0068 | 0.0202 | 0.0068 | −0.0098 |
| | 1 | 0.0187 | 0.9777 | 0.0175 | −0.0223 | −0.0642 |
| | 2 | 0.0103 | 0.9738 | 0.0100 | −0.0262 | −0.0291 |
| | 3 | 0.0069 | 0.9823 | 0.0067 | −0.0177 | −0.0290 |